e-ISBN 978-623-93343-3-8
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 April 2021

PREDIKSI PROGRAM STUDI CALON MAHASISWA UPN VETERAN JAKARTA PADA RUMPUN IPA MENGGUNAKAN *K-NEAREST NEIGHBORS*

Muhammad Labib Mu'tashim, Bagas Aditya Wibisono, Matthew Richard Arianto, Desta Sandya Prasvita, S.Komp., M.Kom.
Program Studi Informatika/ S1 Fakultas Ilmu Komputer
Universitas Pembangunan Nasional Veteran Jakarta
Jl. RS. Fatmawati Raya, Pd. Labu, Kec. Cilandak, Kota Depok, Daerah Khusus Ibukota Jakarta 12450
muhammadlm@upnvj.ac.id, bagasaw@upnvj.ac.id, matthewra@upnvj.ac.id, desta.sandya@upnvj.ac.id

Universitas Pembangunan Nasional Veteran Jakarta atau yang disingkat sebagai UPNVJ adalah salah satu dari banyaknya PTN yang ada di Jakarta membuat banyaknya peminat dan persaingan pada setiap program studinya. Atas dasar itulah penelitian ini bertujuan untuk melihat seberapa besar prediksi program studi rumpun IPA yang dipilih calon mahasiswa UPN Veteran Jakarta menggunakan teknik Data Mining. Teknik ini melihat nilai mereka dan memprediksikan seberapa besar kemungkinan mereka masuk program studi yang sudah dipilih sebelumnya. Prediksi nilai ini menggunakan nilai UTBK dan menggunakan metode data mining yaitu *K-Nearest Neighbors. K-Nearest Neighbors* mengklasifikasikan suatu objek berdasarkan data training. Dari data yang sudah diperoleh dari Kaggle.com ini didapat bahwa akurasi tertinggi sebesar 85% dari nilai K =11 dan K=15 yang sudah diuji.

Keyword **:** Data Mining, *K-Nearest Neighbors*, Prediksi, Program Studi

1. Pendahuluan

Seleksi Bersama Masuk Perguruan Tinggi Negeri atau yang biasa disebut SBMPTN adalah satu jalur dari banyaknya jalur yang tersedia untuk menempuh perguruan tinggi negeri yang dilaksanakan bersamaan oleh para siswa di seluruh Indonesia. SBMPTN sebagai ajang untuk memilih jurusan yang diinginkan para siswa, namun acapkali banyak siswa yang hanya memilih berdasarkan universitas yang terkenal dan program studi yang memiliki keketatan cukup tinggi. Dengan nilai UTBK yang sudah di dapat dan juga dari banyaknya siswa yang memilih suatu jurusan dari banyaknya jurusan yang ada, maka persentase kemungkinan masuk pun akan dipertanyakan.

UPNVJ adalah salah satu universitas negeri yang berada di Jakarta, Indonesia. Universitas ini berdiri pada 21 Februari 1967 hingga sekarang. Selain di Jakarta, UPN juga memiliki kampus lain seperti UPNVY di Yogyakarta, dan UPNVJT di Jawa Timur. Sebelum menyandang statusnya sebagai universitas negeri, UPN merupakan universitas kedinasan milik Dephankam RI. Kemudian berubah menjadi universitas swasta pada tahun 1994, dan menjadi universitas negeri pada tahun 2014.

Maka dari itu dengan menggunakan nilai UTBK yang ada, dengan bantuan menggunakan salah satu algoritma dalam Data Mining, yaitu *K-Nearest Neighbors*, diharapkan dapat memprediksi seberapa besar persentase siswa masuk ke jurusan yang diinginkan. *K-Nearest Neighbors* adalah algoritma klasifikasi yang cara kerjanya dengan mengambil sejumlah K terdekat atau tetangga untuk menentukan kelas dari data baru. Algoritma ini mengelompokkan berdasarkan kemiripan dengan tetangga lainnya. Jumlah nilai yang mirip antara satu jurusan dengan jurusan lain pada rumpun IPA UPN Veteran Jakarta ini membuat penelitian menggunakan KNN sangatlah cocok untuk diterapkan.

Dari permasalahan diatas, dapat dirumuskan bahwa bagaimana cara memprediksi akurasi siswa yang akan masuk ke Universitas Pembangunan Nasional Veteran Jakarta Rumpun Ilmu Pengetahuan Alam menggunakan Algoritma *K-Nearest Neighbors*.

e-ISBN 978-623-93343-3-8
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 April 2021

2. Landasan Teori

2.1 Data Mining

Data Mining merupakan proses untuk *me-mining* data atau menggali suatu data. Data disini merupakan data yang amat besar untuk mencari informasi penting didalamnya. Proses analisa inilah digunakan untuk memperoleh sesuatu yang baru, bermanfaat yang terkadang tidak disadari keberadaannya.

Data Mining berupaya untuk menemukan suatu pola atau corak suatu data yang memakai teknik dari matematika, statistik, kecerdasan buatan, hingga machine learning untuk mengetahui dan mengidentifikasi suatu informasi dari data yang amat besar. Langkah-langkah prosesnya terdiri dari pre-processing, transformasi, hingga mendapatkan output/ hasil.

Ada beberapa tahapan dalam Data Mining itu sendiri, yaitu dari proses *Pre-Processing*, *Data Transformation*, hingga *Evaluation*. Mengacu pada literature sebelumnya (nomor[5]), data yang sudah diambil harus dibersihkan terlebih dahulu (*cleaning*) dari data yang kurang lengkap/ hilang (*missing value*). Kemudian juga Transformasi Data dengan menggabungkan unsur-unsur agar saling melengkapi dan menjadikan suatu data yang lengkap, lalu *Evaluation Data* untuk menafsirkan hasil yang sudah keluar.

2.2 K-Nearest Neighbors

Algoritma *K-Nearest Neighbors* atau biasa yang disebut dengan K-NN adalah suatu metode klasifikasi dari kumpulan data yang mana hasil dari *instance* yang baru di klasifisikan berdasarkan daripada kedekatan jarak tetangga-tetangga terdekat. KNN ini termasuk kedalam *supervised learning*, yaitu menggunakan data yang sudah ada dan begitu juga dengan outputnya, yang menggunakan *Neighborhood Classification* untuk mendapatkan nilai prediksi yang baru.

Tujuan penggunaan KNN ini adalah untuk mengklasifisikan objek yang baru dari atribut-atribut dan sampel dari data training.

2.3 Python dan Spyder IDE

Salah satu perangkat IDE terbaik untuk Python adalah Spyder, yang mana python sangat cocok jika ingin menghitung data menggunakan metode dari Data Mining. Spyder juga bisa menampilkan grafik maupun statistik yang sangat diperlukan bagi yang menggunakan Data Mining.

3. Metode Penelitian

3.1 Pengumpulan Data

Penelitian ini menggunakan dataset dari *kaggle.com* yang berisi data nilai peserta UTBK tahun 2019 di seluruh Indonesia.

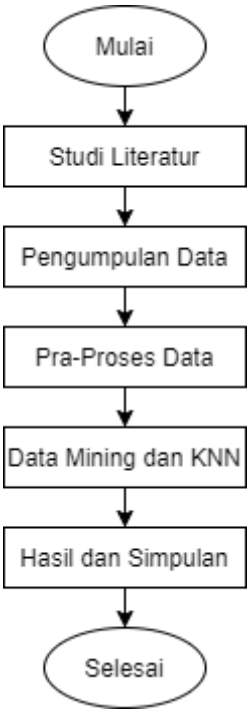
Selanjutnya pada proses penelitian ini dengan mempelajari dan mengumpulkan literatur yang berkaitan erat dengan konsep Data Mining, terutama yang menggunakan Algoritma *K-Nearest Neighbors*. Sumber yang didapat berupa teks dokumen, buku, paper, jurnal, dan situs-situs pendukung untuk mengumpulkan data seperti *kaggle.com*

3.2 Pengolahan Data

Data nilai UTBK 2019 yang didapat dari *kaggle.com* yaitu data peserta UTBK dengan atribut Nomor Urut, Pilihan Prodi 1, Pilihan Prodi 2, Universitas 1, Universitas 2, dan Nilai UTBK masing-masing yaitu Matematika, Biologi, Kimia, Fisika, KPU (Kemampuan Penalaran Umum), Kua (Kemampuan Kuantitatif, PPU (Pengetahuan dan Pemahaman Umum) dan KMB (Kemampuan Memahami Bacaan dan Menulis).

Data-data yang sudah dapat kemudian masuk ke tahap *pre-processing* Data Mining, dimana data yang memiliki atribut kurang lengkap atau kosong *(empty*) akan dihapus. Kemudian tahap selanjutnya adalah *relevation data*, dimana atribut data diseleksi lagi dan hanya mengambil atribut yang memiliki relevansi dengan data yang dibutuhkan. Dalam penelitian ini ada beberapa atribut yang dihapus yaitu Pilihan Prodi 2, Universitas 2, dan juga menghapus pilihan universitas diluar UPN Veteran Jakarta.

Data yang sudah terseleksi akan di lanjutkan ke proses Data Mining dan Algoritma *K-Nearest Neighbors*, disini data akan dibagi menjadi data testing dan data training. Setelah menentukan hal tersebut barulah bisa diketahui prediksi masuk calon mahasiswa UPN Veteran Jakarta Rumpun IPA dengan menggunakan akurasi.



Gambar. 1. Bagan Alur

e-ISBN 978-623-93343-3-8
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 April 2021

4. Pembahasan dan Hasil

4.1 Pra-Proses Data

Di tahap sekarang, data di cleaning agar data-data yang kurang lengkap segera di hapus, karena jika digabung, akan mempengaruhi akurasi akhir.

Tabel. 1. Tabel data awal

df - DataFrame

Index	Jnnamed: (	id_first_major	id_first_university	id_second_major	id_second_university	id_user	score_bio	score_fis	score_kim	score_kmb	score_kpu	score_kua	score_mat	score_ppu
0	0	3321065	332	3331187	333	4	400	400	400	400	400	400	400	400
1	1	3211015	321	3611066	361	14	816	666	651	678	685	706	695	562
2	2	3721093	372	3551302	355	19	562	839	624	700	781	464	551	668
3	3	3321096	332	3551194	355	23	700	669	692	679	692	813	507	573
4	4	5211104	521	5211085	521	28	461	619	441	593	563	500	666	370
5	5	6111014	611	3331021	333	29	516	503	410	717	614	641	479	512
6	6	3411041	341	3411122	341	33	675	584	578	670	647	706	478	590
7	7	3531031	353	3531197	353	37	641	567	600	627	553	597	529	691
8	8	7111196	711	7111092	711	39	440	410	554	645	587	502	649	447
9	9	3211015	321	7111076	711	43	438	459	532	524	442	654	535	537
10	10	7111076	711	7111076	711	45	513	501	474	479	445	430	415	465

Tampilan diatas merupakan 10 data teratas dari Tabel Nilai UTBK dari 10 ribu siswa, kemudian, *cleaning data* dan mengurangi atribut yang tidak diperlukan.

Tabel. 2. Tabel data yang sudah di *cleaning*

df - DataFrame

Index	Jnnamed: (	id_first_major	id_first_university	id_user	score_bio	score_fis	score_kim	score_kmb	score_kpu	score_kua	score_mat	score_ppu
113	113	3241045	324	662	756	702	613	667	591	621	447	623
185	185	3241084	324	1029	573	442	585	364	451	526	538	446
206	206	3241053	324	1104	420	513	597	481	316	417	655	521
226	226	3241053	324	1196	482	662	472	538	745	661	602	491
242	242	3241111	324	1260	393	468	435	395	610	676	489	335
283	283	3241045	324	1469	501	507	460	486	555	493	559	616
306	306	3241037	324	1592	489	530	539	585	387	380	745	413
319	319	3241092	324	1665	733	489	547	552	524	473	433	684
339	339	3241014	324	1769	615	575	479	534	587	687	523	485
362	362	3241076	324	1902	499	455	478	616	565	506	476	571

Data yang sudah diseleksi berkurang menjadi 1422 baris dari 86 ribu baris. Pada data ini hanya diambil siswa yang memilih UPN Veteran Jakarta pada Pilihan Prodi 1. Kolom yang dihapus meliputi Pilihan Prodi 2 dan Universitas 2.

e-ISBN 978-623-93343-3-8
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 April 2021

4.2 Transformasi Data

Data yang sudah di *cleaning* belum memiliki banyak atribut yang diperlukan, maka dari itu perlu juga penambahan atribut baru yaitu rata-rata nilai UTBK setiap siswa dan kode program studi yang dipilih siswa pada Pilihan Prodi 1.

Tabel. 3. Tabel Lengkap

df_new - DataFrame

Index	id_user	nilai_bio	nilai_fis	nilai_kim	nilai_kmb	nilai_kpu	nilai_kua	nilai_mat	nilai_ppu	nilai_rerata	prodi_pilih
113	662	756	702	613	667	591	621	447	623	627.5	3241045
185	1029	573	442	585	364	451	526	538	446	490.625	3241084
206	1104	420	513	597	481	316	417	655	521	490	3241053
226	1196	482	662	472	538	745	661	602	491	581.625	3241053
242	1260	393	468	435	395	610	676	489	335	475.125	3241111
283	1469	501	507	460	486	555	493	559	616	522.125	3241045
306	1592	489	530	539	585	387	380	745	413	508.5	3241037
319	1665	733	489	547	552	524	473	433	684	554.375	3241092
339	1769	615	575	479	534	587	687	523	485	560.625	3241014
362	1902	499	455	478	616	565	506	476	571	520.75	3241076

Penjelasan Kode Program Studi :

Tabel. 4. Tabel Kode dan Passing Grade UPNVJ

Kode	Program Studi	Passing Grade/ Batasan
3241014	Teknik Mesin	606,25
3241022	Teknik Perkapalan	599,16
3241037	Teknik Industri	616,43
3241045	Informatika	626,91
3241053	Sistem Informasi	616,78
3241061	Kedokteran	678,13
3241076	Keperawatan	597,71
3241084	Kesehatan Masyarakat	612,84
3241092	Gizi	611,83
3241103	Teknik Elektro	562,28
3241111	Farmasi	619,05

Sumber PG : https://unjkita.com/jurusan-di-upn-jakarta/

Sumber Kode Prodi : https://sidata-ptn.ltmpt.ac.id/ptn_sb.php?ptn=324

e-ISBN 978-623-93343-3-8
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 April 2021

Tahap selanjutnya adalah normalisasi data menggunakan salah satu metode pada Data Mining, yaitu *Min-Max Normalization*. Pada proses ini, nilai diubah ke rentang nilai 0 sampai 1.

Tabel. 5. Tabel data yang sudah di lakukan normalisasi menggunakan *Min-Max Normalization*

df_scaled - DataFrame

Index	nilai_bio	nilai_fis	nilai_kim	nilai_kmb	nilai_kpu	nilai_kua	nilai_mat	nilai_ppu	nilai_rerata
0	0.788871	0.704545	0.624357	0.682836	0.545966	0.599656	0.231419	0.663534	0.735962
1	0.489362	0.25	0.576329	0.117537	0.283302	0.436426	0.385135	0.330827	0.299881
2	0.238953	0.374126	0.596913	0.335821	0.0300188	0.249141	0.58277	0.471805	0.297889
3	0.340426	0.634615	0.382504	0.442164	0.834897	0.668385	0.493243	0.415414	0.589805
4	0.194763	0.295455	0.319039	0.175373	0.581614	0.694158	0.302365	0.12218	0.250498
5	0.371522	0.363636	0.361921	0.345149	0.478424	0.379725	0.420608	0.650376	0.400239
6	0.351882	0.403846	0.497427	0.529851	0.163227	0.185567	0.734797	0.268797	0.35683
7	0.751227	0.332168	0.511149	0.468284	0.420263	0.345361	0.20777	0.778195	0.502987
8	0.558101	0.482517	0.394511	0.434701	0.538462	0.713058	0.359797	0.404135	0.522899
9	0.368249	0.272727	0.392796	0.587687	0.497186	0.402062	0.280405	0.565789	0.395858
10	0.538462	0.309441	0.439108	0.445896	0.450281	0.247423	0.320946	0.287594	0.330944

4.3 Pembagian Data

Data yang sudah dinormalisasi selanjutnya masuk ke tahap pembagian data. Data dibagi menjadi data uji atau testing dan data latih atau training. Data yang sudah di *cleaning*, didapat total datanya ada 1422 baris. Pembagian data ini meliputi 70% data training dan 30% data testing. Penelitian ini menggunakan Data Training sebesar 70% karena total data yang sudah di normalisasi sekitar 1000 lebih data, dan itu adalah data yang cukup banyak, maka lebih efektif menggunakan proporsi 70 : 30 dibanding 80 : 20. Dari pembagian data tersebut, data training sebanyak 995 data dan data testing sebanyak 427 data.

4.4 Hasil

Data latih dan data uji yang sudah ada kemudian dilanjutkan dengan metode *K-Nearest Neighbors*, dimana pada tahap pemodelan ini ditentukan dari banyaknya K. pada penelitian ini melakukan proses percobaan sebanyak 8 K untuk melihat seberapa besar akurasi dari data testing yang sebanyak 427 data. Dari delapan K akan dipilih akurasi yang paling tinggi.

Tabel. 6. Tabel Akurasi K-NN

Nomor	Nilai K	Akurasi
1	2	0.7939110070257611
2	3	0.8126463700234192
3	5	0.8266978922716628
4	7	0.8477751756440282
5	9	0.8454332552693209
6	11	0.8594847775175644
7	13	0.8477751756440282
8	15	0.8524590163934426

Dari pengujian diatas menggunakan data testing dengan nilai K yang berbeda membuat nilai akurasi nya pun berbeda. Dilihat dari tabel, akurasi terbesar didapatkan dari beberapa nilai K = 11 dan K = 15 sebesar 85%. Lain hal dengan nilai K = 9 dengan akurasi 0,84 atau 84%, dan akurasi terkecil diperoleh pada nilai K = 2 sebesar 0,79 atau 79%. Dari perbandingan ini, bisa disimpulkan bahwa hasil prosentase tidak ditentukan spesifik dari salah

satu nilai K, karena setiap nilai K yang diuji mendapatkan hasil yang cukup dekat satu sama lain, dan bisa disimpulkan akurasi yang didapat berkisar 84 – 85%.

5. Kesimpulan

Penelitian ini menggunakan *K-Nearest Neighbors* untuk menentukan seberapa besar akurasi atau keberhasilan metode yang digunakan untuk memprediksikan calon mahasiswa yang memilih program studi rumpun IPA di UPN Veteran Jakarta. Pengujian data menggunakan metode *K-Nearest Neighbors*, *Min-Max Normalization*, lalu juga pembagian data training dan data testing. Pengujian data dengan nilai K=2 mendapatkan akurasi sebesar 79%, lalu nilai K=3 mendapatkan hasil akurasi 81%, kemudian nilai K=5 mendapatkan akurasi sebesar 82%, lalu pada nilai K=7, K=9, dan K=13 didapat akurasi 84%, sedangkan untuk K=11 dan K=15 mendapatkan akurasi terbesar yaitu sebesar 85%.

Dari kesimpulan diatas bisa didapat bahwa nilai akurasi semakin meningkat seiring tingginya nilai K itu tidak sepenuhnya benar, melihat pada percobaan kali ini setiap hasil dari nilai K memiliki prosentasi yang mirip satu sama lain. Maka dari itu bisa disimpulkan bahwa akurasi tertinggi sebesar 85% didapat dengan dari nilai K yaitu 11 dan 15.

Untuk penelitian selanjutnya diharapkan dapat mencari data dan atribut yang lebih banyak sehingga akurasinya pun semakin berkualitas dan semakin mendekati output yang diinginkan, dan juga bisa menggunakan metode Data Mining lainnya untuk membandingkan output dan akurasi yang didapat.

6. Daftar Pustaka

[1] Arum Sari, Citra dkk.2014. Kluster K-Means Data Mahasiswa Baru Terhadap Program Studi Yang Dipilih, Nomor 136-143

[2] Ashari. 2012. Aplikasi Data Mining Untuk Memprediksi Mata Kuliah Pilihan Pada Program Studi Teknik Informatika STMIK AKBA, Nomor 29-36

[3] Asril, Elvira dkk.2015. Analisis Data Lulusan dengan Data Mining untuk Mendukung Strategi Promosi Universitas Lancang Kuning. *Jurnal Teknologi Informasi & Komunikasi Digital Zone,* Volume 6, Nomor 2, November 2015: 24-32

[4] Rahmawati, Arindiah dkk.2018. Prediksi Penentuan Program Studi Menggunakan Algoritma K-Nn Pada Pelajar SMAN 6 Kota Depok Jurusan Ipa. *Seminar Nasional Informatika, Sistem Informasi Dan Keamanan Siber (SEINASI-KESI)* Jakarta-Indonesia, 1 Desember 2018, Nomor 192-197

[5] Sumpena dkk. 2018. Penerimaan Calon Siswabaru dan Penentuan Penjurusan dengan Algoritma C 4.5 SMK Plus PGRI 1 Cibinong, *CKI On SPOT*, Vol. 11, No. 2

[6] Swastina, Liliana. 2013. Penerapan Algoritma C4.5 Untuk Penentuan Jurusan Mahasiswa, *Jurnal GEMA AKTUALITA*, Vol. 2 No. 1

[7] https://www.kaggle.com/ekojsalim/indonesia-college-entrance-examination-utbk-2019 (diakses 2 April 2021).